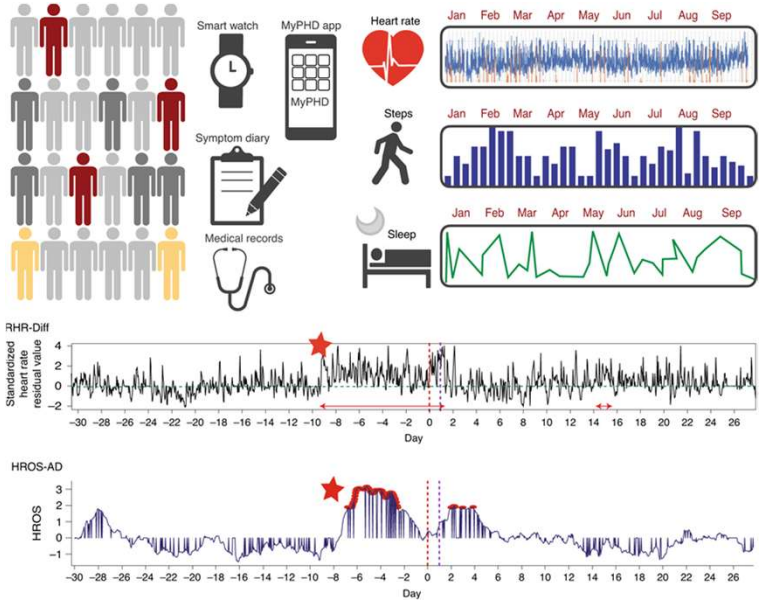
Universidad Popular del Cesar

Minería de datos y Patrones

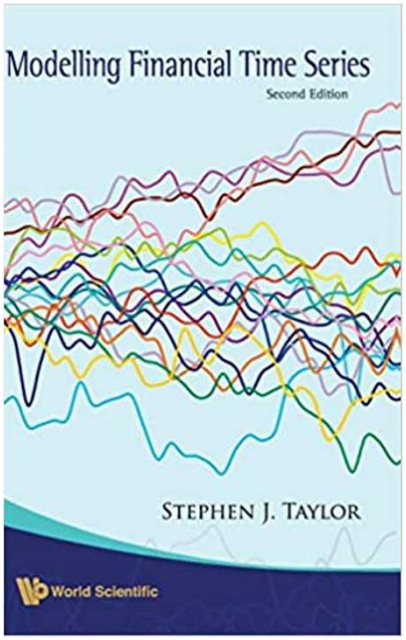
Version 2025-I

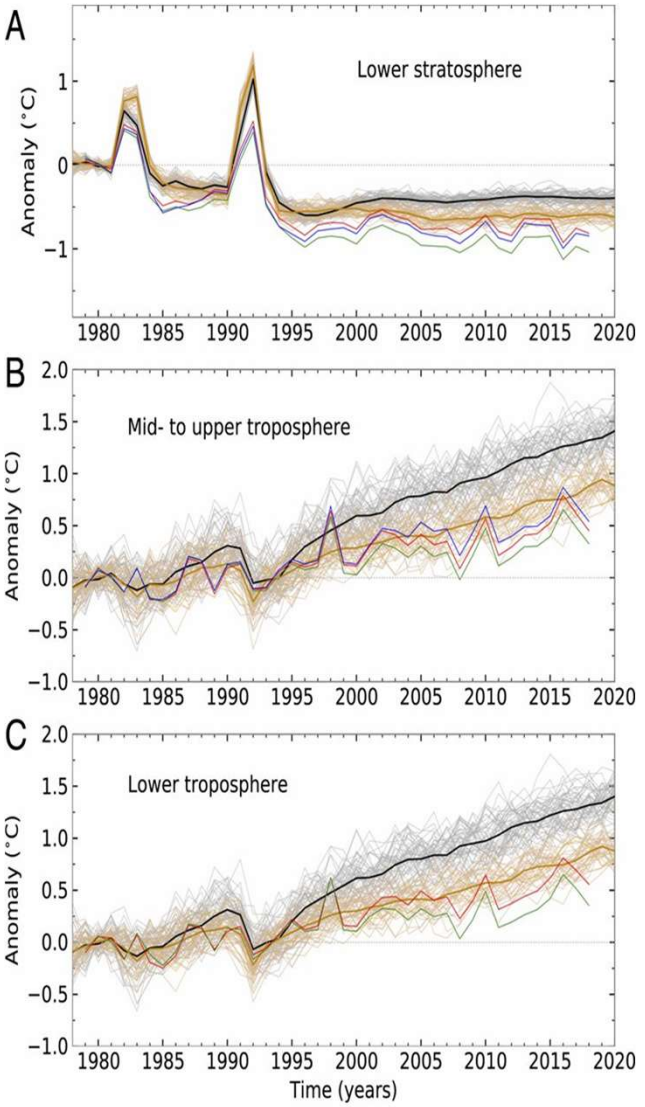
Clustering de Series de Tiempo

Dr. José Ramón Iglesias
DSP-ASIC BUILDER GROUP
Director Semillero TRIAC
Ingenieria Electronica
Universidad Popular del Cesar



Mishra, T.. *et al.* Pre-symptomatic detection of COVID-19 from smartwatch data. *Nat Biomed Eng* (2020).





Santer, Benjamin D., et al. "Quantifying stochastic uncertainty in detection time of human-caused climate signals." *PNAS* (2019)

Clustering basado en la forma usando el algoritmo K-means

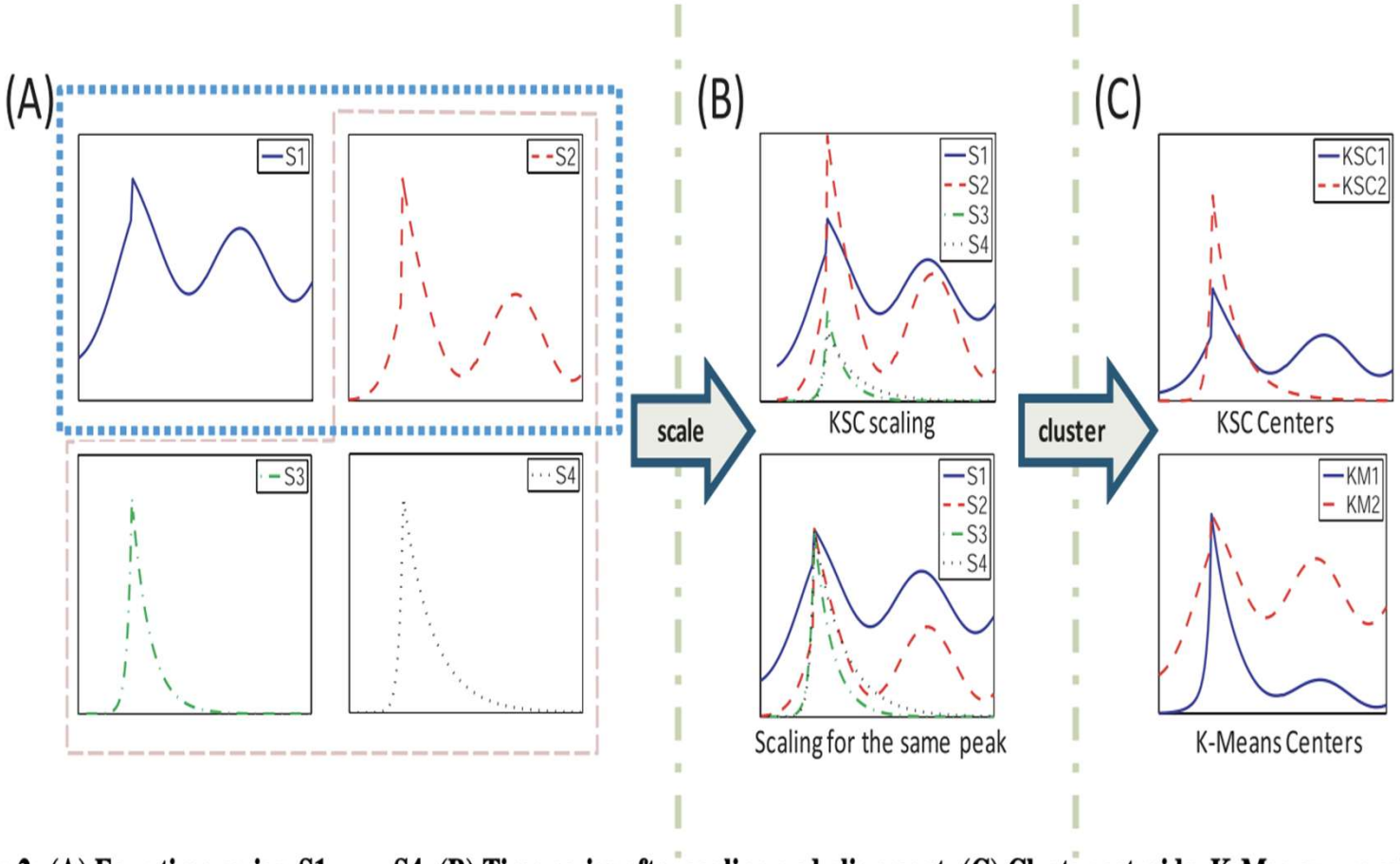


Figure 2: (A) Four time series, S1, . . . , S4. (B) Time series after scaling and alignment. (C) Cluster cetroids. K-Means wrongly puts {S1} in its own cluster and {S2, S3, S4} in the second cluster, while *K-SC* nicely identifies clusters of two vs. single peaked time series.

Yang, Jaewon, and Jure Leskovec. "Patterns of temporal variation in online media." *ACM international conference on Web search and data mining*. 2011

Medida de distancia

$$\hat{d}(x, y) = \min_{\alpha, q} \frac{||x - \alpha y_{(q)}||}{||x||}$$

$$F = \sum_{k=1}^{K} \sum_{x_i \in C_k} \hat{d}(x_i, \mu_k)^2.$$

$$\mu_k^* = \arg\min_{\mu} \sum_{x_i \in C_k} \hat{d}(x_i, \mu)^2.$$

Yang, Jaewon, and Jure Leskovec. "Patterns of temporal variation in online media." *ACM international conference on Web search and data mining*. 2011

Cálculo de Centroide

$$\mu_k^* = \arg\min_{\mu} \sum_{x_i \in C_k} \min_{\alpha_i, q_i} \frac{\left|\left|\alpha_i x_{i(q_i)} - \mu\right|\right|^2}{\left|\left|\mu\right|\right|^2}$$

Finally, substituting $\sum_{x_i \in C_k} (I - \frac{x_i x_i^T}{||x_i||^2})$ by M leads to the following minimization problem:

$$\mu_k^* = \arg\min_{\mu} \frac{\mu^T M \mu}{||\mu||^2}. \quad (4)$$

Resultado de álgebra lineal: La solución de este problema es el vector propio (eigenvector) u correspondiente al valor propio más pequeño λ de la matriz M

Extensión: Series de tiempo multidimensionales

Algorithm 1 m-kSC Algorithm

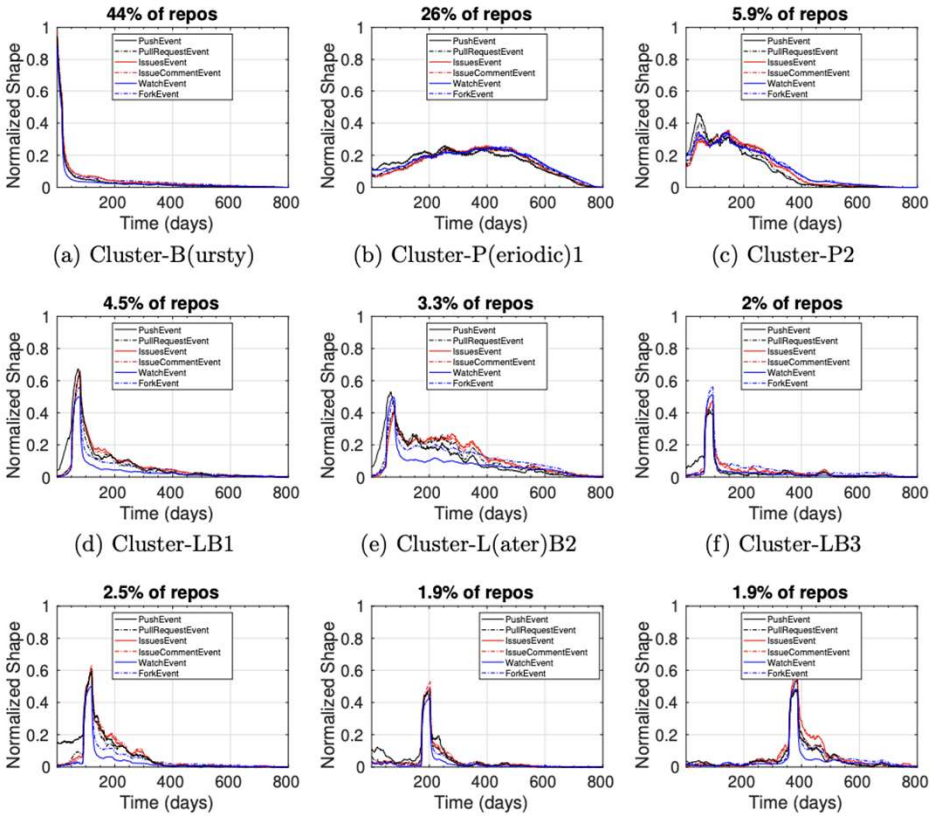
Input:$\{\mathcal{X}, K\}$ where $\mathcal{X} \in \mathbb{R}^{N \times D \times M}$ is the tensor containing N multidimensional time series and K is number of clusters.

Output: $\{\mathcal{C}, S\}$ where $\mathcal{C} \in \mathbb{R}^{K \times D \times M}$ is the tensor of cluster centroids and S contains each cluster assignments.

Initialize cluster assignments S randomly
while S changes on every iteration **do**
  for $k = 1 : K$ **do**
    for $d = 1 : D$ **do**
      $M = \sum_{\mathbf{x}_n \in S_k} (I - \frac{\mathbf{x}_n(d,:)\mathbf{x}_n(d,:)^T}{\|\mathbf{x}_n(d,:)\|^2})$
      $\mathbf{C}(k,d,:) =$ Smallest eigenvector of M.
    end for
  end for
  for $n = 1 : N$ **do**
    $k = \underset{k=1,\ldots,K}{\text{argmin}} dist(\mathbf{c}_k, \mathbf{x}_n)$ using Eq. 1
    $S(n) = k$
  end for
end while

Ozer, M., Sapienza, A., Abeliuk, A., Muric, G., & Ferrara, E. (2020). Discovering patterns of online popularity from time series. *Expert Systems with Applications*.

Patrones temporales de la evolución de los 1.000 repositorios de GitHub más populares.



(a) Cluster-B(ursty) (b) Cluster-P(eriodic)1 (c) Cluster-P2

(d) Cluster-LB1 (e) Cluster-L(ater)B2 (f) Cluster-LB3

(g) Cluster-LLB1 (h) Cluster-LLB2 (i) Cluster-LLB3

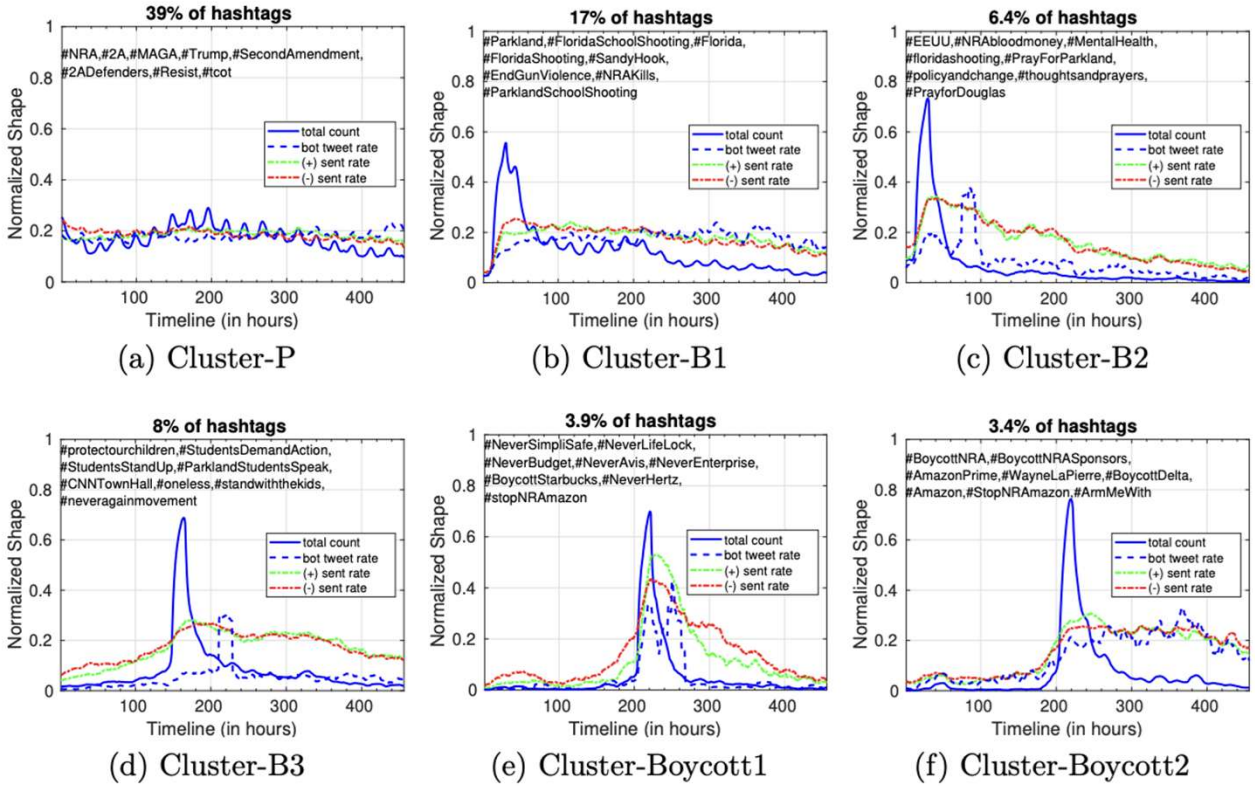
Figure 3: Shapes of the uncovered cluster centroids in the GitHub dataset.

Ozer, M., Sapienza, A., Abeliuk, A., Muric, G., & Ferrara, E. (2020). Discovering patterns of online popularity from time series. *Expert Systems with Applications.*

Patrones de popularidad en Twitter

Análisis de la línea de tiempo de los top mil hashtags más populares de Twitter.

Datos desde el 14 de febrero hasta 6 de marzo 2018, relacionado con el tiroteo en la escuela de Parkland.



(a) Cluster-P (b) Cluster-B1 (c) Cluster-B2

(d) Cluster-B3 (e) Cluster-Boycott1 (f) Cluster-Boycott2

Figure 7: Shape of the uncovered clusters of Twitter

Ozer, M., Sapienza, A., Abeliuk, A., Muric, G., & Ferrara, E. (2020). Discovering patterns of online popularity from time series. *Expert Systems with Applications*.